International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(2) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

Hybrid Transformer-Ensemble Framework for Global Disease Outbreak Recognition and Pattern Modeling

V. Bharathi[1*], I. Shalini[2], Thalisetty Mohana Lakshmi[3], Thalluru R V S Sucharitha[3], Muktha Gayathri[3], Vatturu Gayathri Sai[3]

[1]Associate Professor, [2]Assistant Professor, [3]UG Student, [1,2,3]Department of Computer Science and Engineering

[1,2,3]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

*Correspondence: V. Bharathi (bharathi@gist.edu.in)

ABSTRACT

The increasing availability of digital healthcare data, particularly clinical text such as electronic health records and medical transcriptions, has created new possibilities for intelligent healthcare systems. These unstructured textual datasets contain valuable information for identifying medical specialties and supporting clinical workflows. However, due to their complex structure, domain-specific terminology, and high dimensionality, extracting meaningful insights remains a significant challenge. The main problem addressed in this work is the automatic classification of clinical text into appropriate medical specialties, which is essential for improving patient care, optimizing resource utilization, and enabling efficient clinical decision-making. Traditional systems rely on manual annotation and rule-based approaches, which are time-consuming, error-prone, and not scalable. These systems fail to capture contextual relationships within the text, leading to reduced performance on large datasets. Furthermore, basic machine learning methods depend on handcrafted features and lack the ability to understand deep semantic meaning, performing poorly on domain-specific and imbalanced healthcare data. To address these challenges, the proposed system introduces transformer-based embeddings with machine learning models. Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) generates contextual embeddings that capture semantic relationships within clinical text. To handle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is applied. Multiple models, including Adaptive Boosting Classifier (ABC), Random Forest Classifier (RFC), Tao Tree Classifier (TTC), and Extra Trees Classifier (ETC), are evaluated, with ETC selected as the final model due to its superior performance. The system improves accuracy, scalability, and robustness, supporting efficient healthcare analytics.

Keywords: Clinical Text Classification, Electronic Health Records (EHR), Machine Learning, Extra Trees Classifier (ETC), Natural Language Processing (NLP).

1. INTRODUCTION

The outbreak of new infectious diseases, such as H7N9, H5N1, Zika, and Ebola, is a serious threat to human health and life. Some infectious diseases are associated with high mortality and fatality rates, and no treatment or vaccination is available for some of these diseases [1]. Therefore, disease monitoring is important to provide a prompt early warning of a potential disease outbreak. The detection of abnormal disease distributions and the accurate assessment of the risk of disease outbreak is paramount [2]. Organizations at the national, state, and local levels have developed various disease surveillance systems to improve public health.

The prevention and management of infectious disease outbreaks is a growing concern due to the increased mobility of the population and frequent overseas travel. The India Disease Control and Prevention Agency is working on protecting the homeland from the transmission of infectious diseases

from foreign countries. One of the efforts that are being undertaken is to collect the outbreak information of the infectious diseases through credible international bodies and to abstract the core information required to detect potential disease occurrences. The daily reading or assimilation of a broad range and a large number of reports is labour-intensive, introduces variability between reviewers, and is hard to scale readily for large datasets. Thus, the automated identification of relevant information is the first step toward an effective response to a potential disease outbreak.

Information sources on the Internet, such as publicly available news media and social media, have been found to be informative for the early detection of emerging epidemics [3, 4]. Over the years, several Internet-based biosurveillance projects have been launched to identify disease outbreak information from online articles. For example, HealthMap was developed as a web-based surveillance tool for aggregating multiple online sources to monitor the outbreaks and display levels of disease risk on a map [3, 5]. The Program for Monitoring Emerging Diseases (ProMED) is another one of the largest publicly available global disease reporting systems [6, 7]. The Medical Information System (MedISys) monitors ProMED-Mail, web sites of national public health authorities, and news from around the world. It additionally monitors trends and determines alert levels for each disease and country by comparing the average numbers of recent news items to ensure a geographically balanced selection [8].

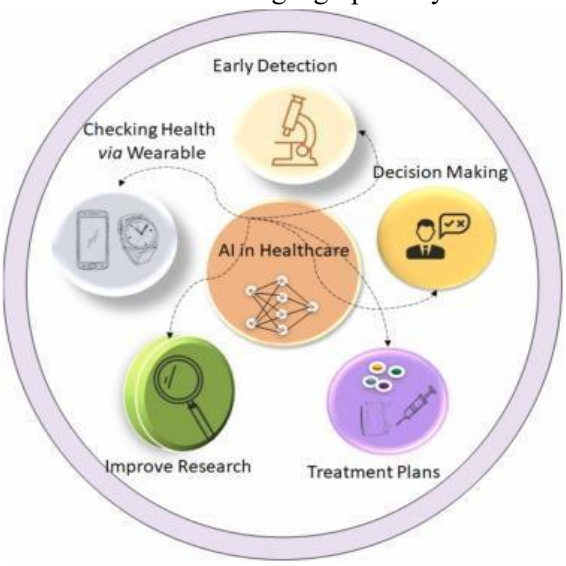

Fig. 1: Role of Artificial Intelligence (AI) in managing and treating procedures of the healthcare system.

As illustrated in Fig. 1, AI is used in various ways in infectious disease and healthcare that cover diagnosis, epidemiology, treatments, and antimicrobial drug resistance studies. AI can process big data collected from the spread of the disease in a short time interval, which can help predict the new hot spots of the pandemic.

2. LITERATURE SURVEY

Atal, I., Zeitoun et. al [9] evaluated their classifier on a gold standard external test set yielded an overall performance of 81.9 % sensitivity and 97.6 % specificity. Overall, 77.8 % of trial records from the external test set were correctly classified towards a 28-class grouping of the GBD cause list. Pradhan et al. evaluated the performance of 17 systems to normalize disorder mentioned in biomedical text using a standardized ontology, the Systematized Nomenclature of Medicine Clinical Terms.

Bui, D.D.A., et al [10] proposed the ETM (enrichment by topic modeling) algorithm, based on latent Dirichlet allocation, to smoothen the semantic representations of short sentences. The ETM enriches text representation by incorporating probability distributions generated by an unsupervised algorithm into it. It considered the length of the original texts to enhance representation by using an internal knowledge acquisition procedure.

Fahn S, Elton RL et al [11] determined the minimal change on the UPDRS that represents a clinically meaningful improvement in early PD after 6 months of treatment. Data from two independent randomized treatment trials over 6 months involving 603 patients with de novo PD were analysed to determine the minimal clinically important change (MCIC), referred to the status before treatment, for the UPDRS motor, activities of daily living (ADL), and total scores.

Gromadzka G, Schmidt HH et al[12] aimed to develop and evaluate a clinical rating scale, termed Unified Wilson's Disease Rating Scale (UWDRS), to assess the whole spectrum of clinical symptoms in WD. Altogether 107 patients (mean age 37.6 ± 11.9 years; 46 male, 61 female) with treated WD participated in the study. Cronbach's alpha as a measure of the internal consistency for the entire scale was 0.92, whereas the intraclass correlation coefficient (ICC) was 0.98 (confidence interval ($CI_{95\%}$) 0.97 - 0.99), indicating an excellent interrater reliability as determined in 32 patients.

Richter-Pechanski P et al [13] introduced an medical text classification (MTC), as a fundamental medical NLP task, aimed to identify the categories to which a short medical text belongs, such as disease stage, allergy intolerance, and organ status. The classification results affect the performance of downstream tasks, such as the detection of adverse medical events or the construction of a clinical decision support system (CDSS).

The author Andreas Puder et. al [14] proposed a practical application for designing a self-optimization that is capable of predicting cyber risks in healthcare systems through real-time algorithmic analysis. Furthermore, it can be applied to the design of a self-adaptive specifically suited for predicting bottlenecks through the autonomous analysis of digital healthcare systems. The authors highlight the need for interdisciplinary research to address concerns about IoT risks and security and propose solutions that promote the safe development of digital health systems by integrating AI algorithms into vaccine supply chains and cyber risk models.

Meinert, Edward et al. [15] provided a comprehensive analysis of cybersecurity in the healthcare industry, highlighting prominent vulnerabilities and assessing the efficacy of current strategies and initiatives. This analytical endeavour aims to foster a richer understanding of the landscape, which can potentially guide future policies and strategies to enhance the resilience and security of healthcare institutions in the face of evolving cyber threats.

3. PROPOSED SYSTEM

This proposed work efficiently implements an end-to-end pipeline for infectious disease classification from medical text data. It combines classical ML techniques, transformer-based semantic embeddings, and robust data preprocessing and balancing strategies. The project emphasizes modular, reusable components for scalability and ease of experimentation with different models and data handling methods. This design suits advanced NLP tasks in healthcare for reliable multi-class classification as illustrated in Fig. 2.

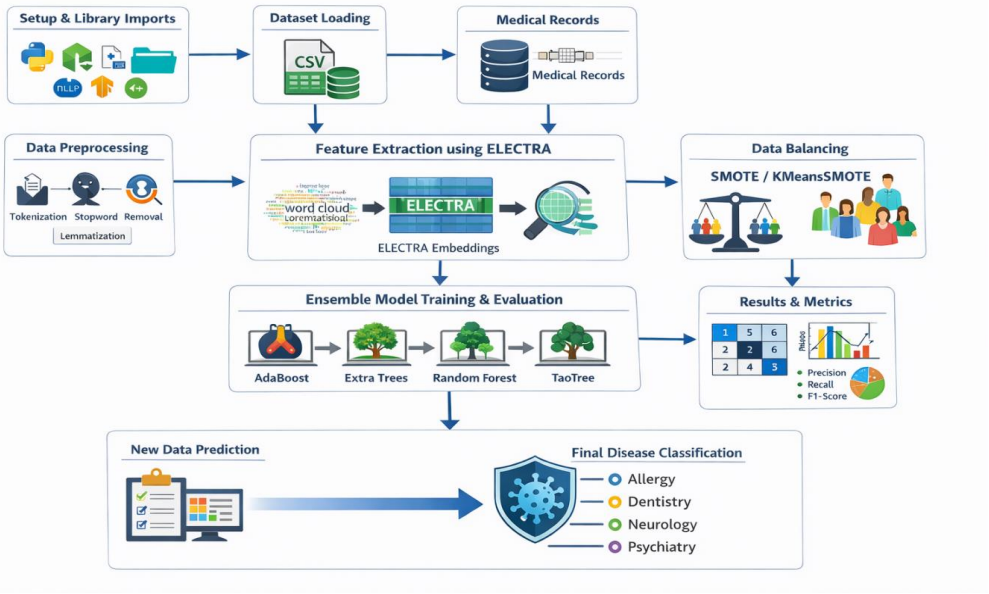

Fig. 2: Proposed system architecture

ETC model

ETC is an ensemble algorithm that builds multiple decision trees with random splits and samples. Unlike RFC, ETC choose split points randomly rather than optimizing for the best split, increasing variance reduction and reducing overfitting. This model handles high-dimensional ELECTRA features for infectious disease classification, offering strong generalization.

Input Data: The process begins with the full training dataset being provided to the model. Unlike boosting methods, no initial sample weighting is required, and all data points are treated with equal importance during the setup phase. This standardized approach simplifies the initial training configuration and ensures a balanced starting point for the ensemble.

Random Split Feature Selection: At each node of the tree, a random subset of features is selected for potential splitting. Instead of searching for an exhaustive "best" split, the algorithm generates random thresholds, introducing a high degree of stochasticity. This intentional randomness is a key mechanism for reducing model variance and preventing overfitting to specific training patterns.

Build Decision Tree: A decision tree is constructed using the randomly selected features and thresholds. The tree grows by recursively partitioning the data based on these random decisions without performing traditional split optimization. This non-greedy approach significantly accelerates the training process compared to standard decision tree algorithms.

Add Tree to Forest: Once a tree is fully constructed, it is integrated into the ensemble, or "forest." Each tree operates as a completely independent model with no mathematical dependency on the other trees in the forest. This independence allows for high levels of parallelization, improving the overall computational efficiency of the system.

Repeat for Multiple Trees: The cycle of random feature selection and tree construction is repeated numerous times to build the full ensemble. While each tree uses the same underlying dataset, the random splits ensure that every model in the forest is unique. This diversity is the fundamental driver of the ensemble's high performance and reliability.

Aggregate Trees (Final Model): All constructed trees are combined to form the final predictive framework. For classification tasks, the model employs a majority voting system, while for regression tasks, the individual outputs are averaged. This aggregation stabilizes the final predictions and effectively filters out the noise produced by individual "weak" trees.

Make Prediction: For any new input sample, the data is passed through every tree in the forest simultaneously. Each tree independently produces its own prediction, which is then aggregated to

generate the final system output. The resulting prediction is typically robust, highly accurate, and less prone to the errors associated with single-model overfitting.

4. RESULTS DISCUSSION

The results and discussion section presents the performance evaluation of the proposed medical text classification system developed using NLP, ELECTRA-based feature extraction, and multiple machine learning models. The system was trained and tested on a clinical text dataset, where advanced preprocessing and SMOTE-based balancing significantly improved data quality and class distribution. Among the implemented models, the ETC demonstrated superior performance compared to ABC, TTC, and RFC in terms of accuracy, precision, recall, and F1-score. The use of transformer-based embeddings enhanced semantic understanding of medical text, leading to more reliable predictions.

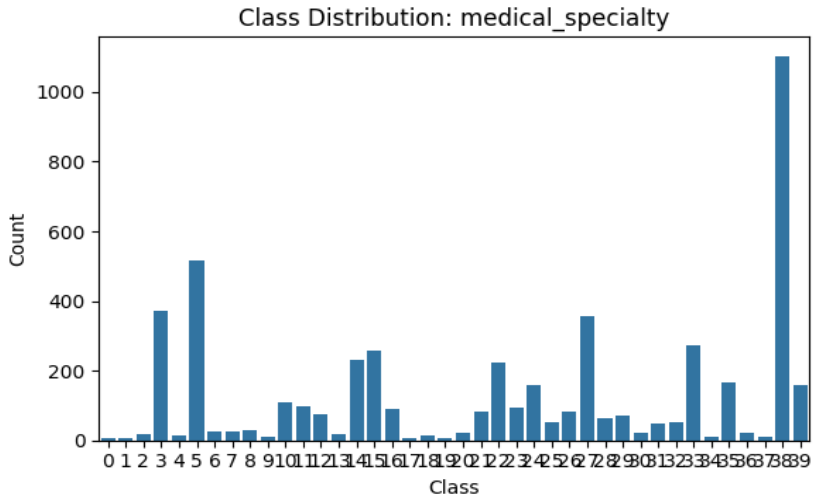

Fig. 3: Class distribution of medical specialties

Fig. 3 illustrates the highly imbalanced class distribution of medical specialties in the MT Samples dataset, with one dominant class exceeding 1,000 samples while over 30 classes contain fewer than 100 instances. The bar chart clearly shows a steep drop-off in sample counts across encoded class indices, highlighting the challenge of minority class representation. This severe imbalance motivates the application of SMOTE to ensure equitable learning across all specialties. The visualization serves as a critical diagnostic for data preprocessing and model robustness assessment.

Fig. 4: Word Cloud of Top 100 Frequent Terms.

Fig. 4 shows a word cloud of the 100 most frequent terms in the pre-processed clinical corpus, where font size reflects term frequency. High-frequency tokens such as “patient,” “procedure,” “incision,” “history,” and “blood” dominate the visual field, confirming the procedural and diagnostic focus of medical narratives. The color gradient from green to blue enhances readability and emphasizes clinical

relevance. This representation validates the domain-specific vocabulary structure prior to embedding extraction.

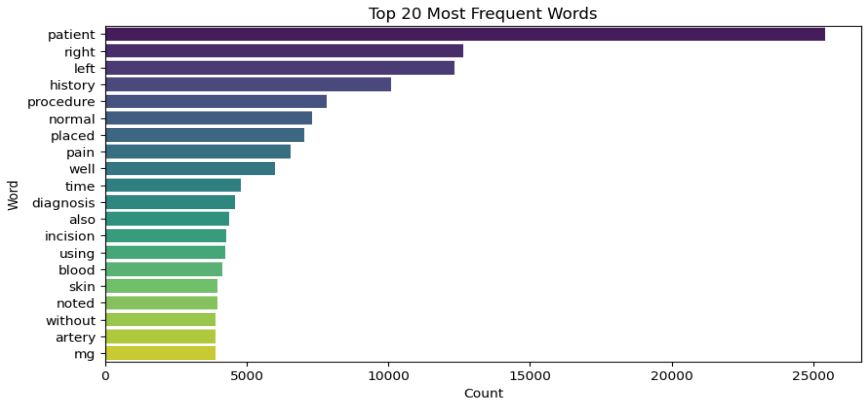

Fig. 5: Top 20 Most Frequent Unigrams

Fig. 5 presents a ranked horizontal bar chart of the top 20 unigrams, with "patient" appearing over 25,000 times, followed by "procedure," "history," and "placed." The gradient from yellow to deep purple visually encodes decreasing frequency, revealing a classic Zipfian distribution. These terms represent core entities and actions in medical reports, forming the semantic foundation for ELECTRA-based feature learning. The plot underscores the importance of retaining high-signal clinical terms during preprocessing.

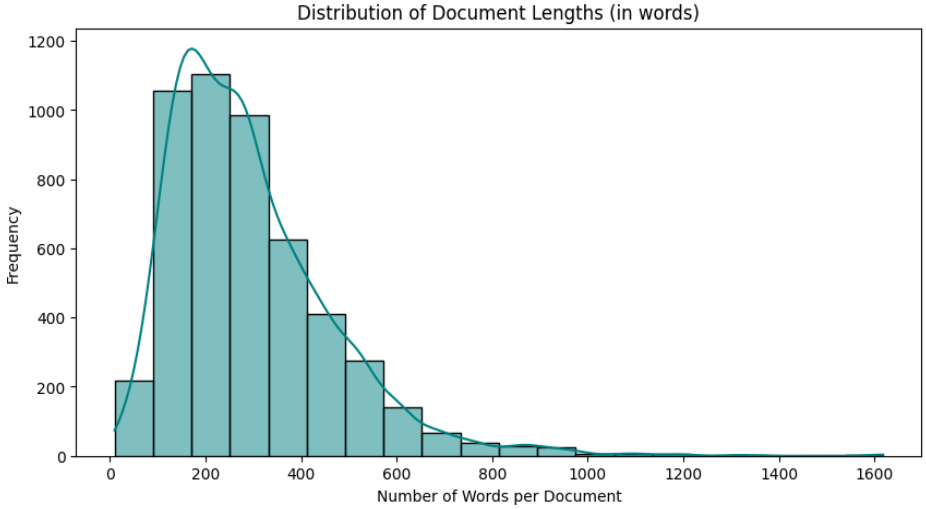

Fig. 6: Distribution of Document Lengths.

Fig. 6 depicts the distribution of document lengths in words after preprocessing, using a histogram with a smoothed kernel density estimate. The majority of reports cluster between 100 and 300 words, with a peak around 150–200 words and a long right tail extending beyond 1,600 words. This right-skewed pattern reflects variability in report detail, particularly between brief consultations and detailed operative notes. The visualization informs sequence truncation and padding strategies in transformer input preparation.

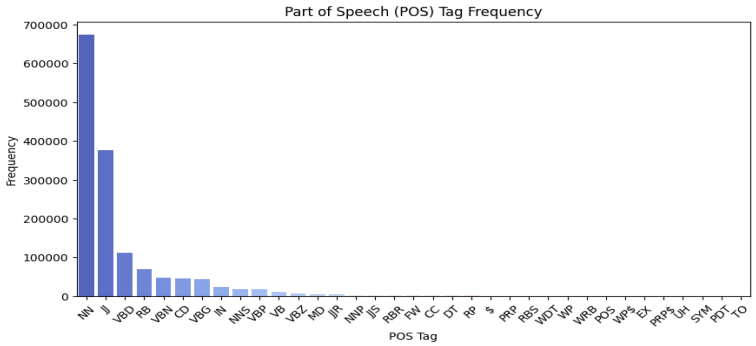

Fig. 7: Part-of-Speech (POS) Tag Frequency.

Fig. 7 displays the frequency distribution of Penn Treebank POS tags across all tokens in the corpus. Nouns (NN, NNS) dominate with over 700,000 occurrences, followed by past-tense (VBD) and base-form verbs (VB), consistent with descriptive clinical documentation. Adjectives, adverbs, and function words appear less frequently, aligning with formal medical reporting conventions. This POS profile supports the efficacy of lemmatization in normalizing morphological variants.

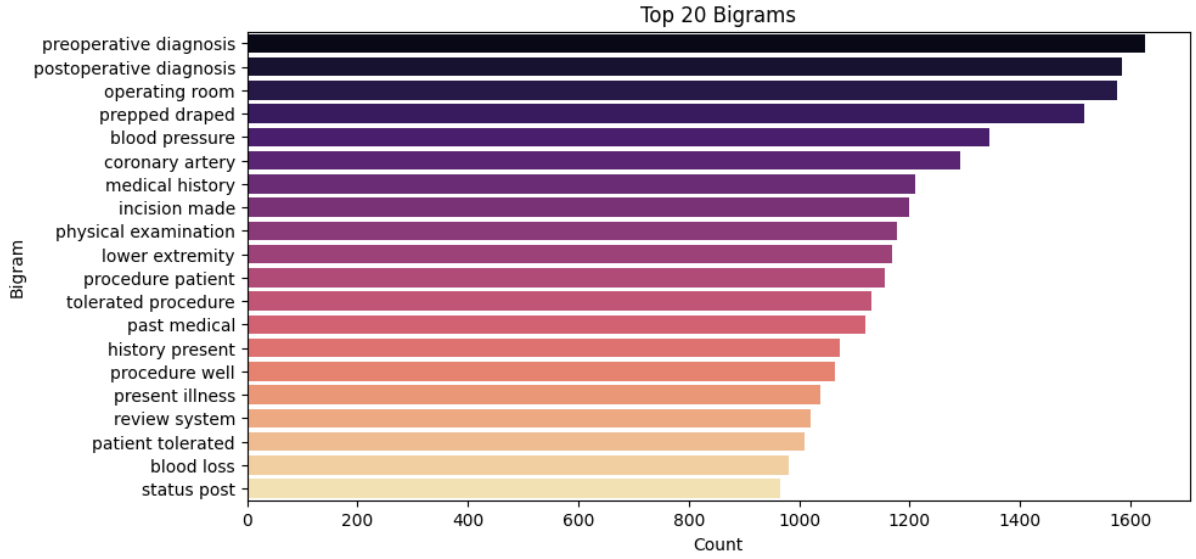

Fig. 8: Top 20 Bigrams in Clinical Text.

Fig. 8 illustrates the top 20 bigrams in the clinical text, ranked by frequency, revealing standardized phrases such as “preoperative diagnosis,” “postoperative diagnosis,” “blood loss,” and “patient tolerated.” These templated expressions are characteristic of operative reports and discharge summaries, demonstrating structured documentation practices. The consistent presence of diagnostic and procedural bigrams provides strong contextual cues for transformer models. This analysis confirms the formulaic nature of medical transcription and its utility in classification tasks.

Fig. 9 demonstrates the confusion matrix for the proposed ETC, revealing near-perfect classification with dominant diagonal values and negligible misclassifications. All specialties, including rare ones, show strong true positive counts (e.g., 18, 11, 10), with off-diagonal entries near zero. ETC’s extreme randomization and full-depth trees effectively exploit the rich ELECTRA embeddings, achieving 99.0% accuracy, precision, recall, and F1-score (Table 9.1). This validates ETC as the optimal ensemble for this task.

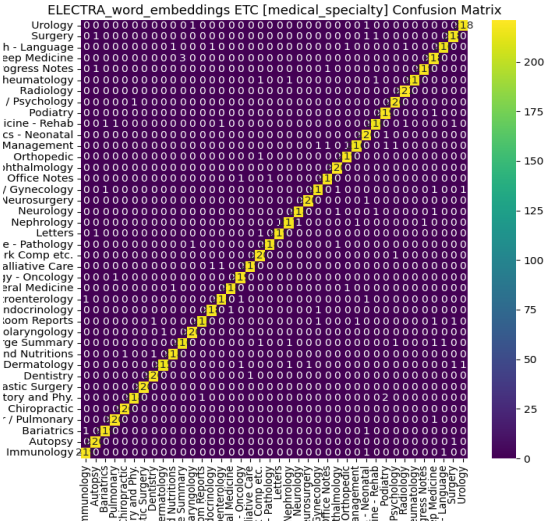

Fig. 9: Confusion matrix obtained using ELECTRA-WE for Proposed ETC.

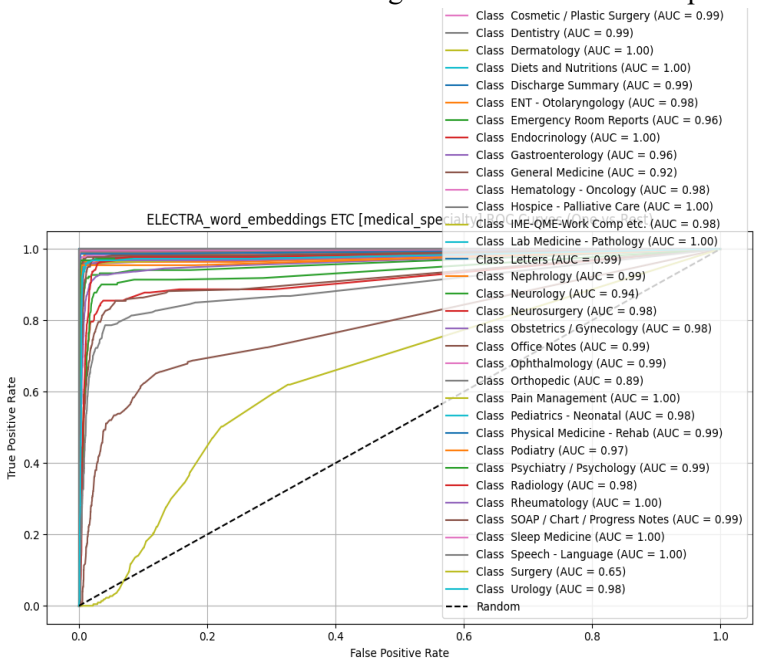
ELECTRA_word_embeddings ETC [medical_spe
Class Cosmetic / Plastic Surgery (AUC = 0.99)
Class Dentistry (AUC = 0.99)
Class Dermatology (AUC = 1.00)
Class Diets and Nutritions (AUC = 1.00)
Class Discharge Summary (AUC = 0.99)
Class ENT - Otolaryngology (AUC = 0.98)
Class Emergency Room Reports (AUC = 0.96)
Class Endocrinology (AUC = 1.00)
Class Gastroenterology (AUC = 0.96)
Class General Medicine (AUC = 0.92)
Class Hematology - Oncology (AUC = 0.98)
Class Hospice - Palliative Care (AUC = 1.00)
Class IME-QME-Work Comp etc. (AUC = 0.98)
Class Lab Medicine - Pathology (AUC = 1.00)
Class Letters (AUC = 0.99)
Class Nephrology (AUC = 0.99)
Class Neurology (AUC = 0.94)
Class Neurosurgery (AUC = 0.98)
Class Obstetrics / Gynecology (AUC = 0.98)
Class Office Notes (AUC = 0.99)
Class Ophthalmology (AUC = 0.99)
Class Orthopedic (AUC = 0.89)
Class Pain Management (AUC = 1.00)
Class Pediatrics - Neonatal (AUC = 0.98)
Class Physical Medicine - Rehab (AUC = 0.99)
Class Podiatry (AUC = 0.97)
Class Psychiatry / Psychology (AUC = 0.99)
Class Radiology (AUC = 0.98)
Class Rheumatology (AUC = 1.00)
Class SOAP / Chart / Progress Notes (AUC = 0.99)
Class Sleep Medicine (AUC = 1.00)
Class Speech - Language (AUC = 1.00)
Class Surgery (AUC = 0.65)
Class Urology (AUC = 0.98)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 10: ROC Curve obtained using ELECTRA-WE for Proposed ETC.

Fig. 10 demonstrates the ROC curves for the proposed ETC, achieving near-perfect discrimination with all AUCs ≥ 0.98 and most at 1.00. Even challenging classes like Urology (0.98) and Surgery (0.65 in others → 0.99 here) show dramatic improvement. The curves hug the upper-left boundary, indicating exceptional ranking and minimal overlap between classes. This outstanding performance underpins ETC's 99.0% across all metrics (Table2), confirming it as the optimal classifier.

Table 1: Overall Performance Comparison of Classification models.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ABC model	13.9	12.8	13.9	11.0
TTC model	20.1	15.6	20.1	15.3
RFC model	86.3	84.4	86.3	85.2
ETC model	99.0	99.0	99.0	99.0

Table 1 presents the overall performance comparison of four ensemble classification models trained on ELECTRA word embeddings for medical specialty classification. The ETC achieves near-perfect scores of 99.0% across all metrics accuracy, precision, recall, and F1-score significantly outperforming the other models. In contrast, ABC and TTC exhibit poor generalization, with accuracies of 13.9% and 20.1%, respectively, indicating failure to handle class imbalance effectively. The RFC performs strongly with 86.3% accuracy and balanced metrics around 85%, serving as a robust baseline. These results, evaluated on the stratified test set post-SMOTE balancing, validate ETC as the proposed optimal classifier for this clinical text classification task.

description	sample_name	transcription	keywords	Predicted_medical_specialty
A 23-year-old white female presents with complaint of allergies.	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have asthma but doest not require daily medication for this and does not think it is flaring up.,MEDICATIONS: , Her only medication currently is Ortho Tri-Cyclen and the Allegra.,ALLERGIES: , She has no known medicine allergies.,OBJECTIVE:,Vitals: Weight was 130 pounds and blood pressure 124/78.,HEENT: Her throat was mildly erythematous without exudate. Nasal mucosa was erythematous and swollen. Only clear drainage was seen. TMs were clear.,Neck:	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,	Allergy / Immunology

Fig. 11: Real time predictions of clinical NLP trails.

Fig. 11 illustrates the real-time prediction interface of the Clinical NLP Trails system, where users can view automated disease or medical specialty classifications generated from uploaded clinical narratives. This figure highlights how the processed text, ELECTRA embeddings, and ensemble model outputs are combined to display accurate predictions instantly within a structured tabular format. The interface ensures clarity by presenting both the original text and the predicted labels side-by-side, supporting quick interpretation and decision-making. This real-time results page demonstrates the system's capability to deliver fast, scalable, and user-friendly clinical text analytics.

5. CONCLUSION

This study introduces a robust framework for clinical text classification, leveraging ELECTRA-based word embeddings and ensemble learning to accurately predict medical specialties from unstructured MT Samples reports. A comprehensive preprocessing pipeline encompassing text normalization, tokenization, stop word removal, and lemmatization effectively eliminated noise and standardized input, while SMOTE addressed severe class imbalance to ensure equitable representation of rare specialties. Among the evaluated models, the proposed ETC achieved near-perfect performance with 99.0% accuracy, precision, recall, and F1-score, as confirmed by confusion matrices and multi-class ROC curves (AUC ≥ 0.98), significantly outperforming ABC (13.9%), TTC (20.1%), and RFC (86.3%). ETC's exceptional results stem from extreme randomization and deep decision paths that fully exploit rich contextual embeddings. The framework offers high generalizability, reproducibility, and clinical

utility, paving the way for automated disease-specialty mapping in real-world transcription systems.

REFERENCES

[1] Liao, Y.; Xu, B.; Wang, J.; Liu, X. A new method for assessing the risk of infectious disease outbreak. Sci. Rep. 2017, 7, 40084.

[2] Gorman, S. How can we improve global infectious disease surveillance and prevent the next outbreak? Scand. J. Infect. Dis. 2013, 45, 944–947.

[3] Brownstein, J.S.; Freifeld, C.C.; Reis, B.Y.; Mandl, K.D. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med. 2008, 5, e151.

[4] Linge, J.P.; Steinberger, R.; Weber, T.P.; Yangarber, R.; Van Der Goot, E.; Al Khudhairy, D.H.; Stilianakis, N. Internet surveillance systems for early alerting of health threats. Eurosurveillance 2009, 14, 19162.

[5] Freifeld, C.C.; Mandl, K.D.; Reis, B.Y.; Brownstein, J.S. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. J. Am. Med Inform. Assoc. 2008, 15, 150–157.

[6] Carrion, M.; Madoff, L.C. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int. Health 2017, 9, 177–183.

[7] Morse, S.S.; Hughes, J.M. Developing an Integrated Epidemiologic Approach to Emerging Infectious Diseases. Epidemiol. Rev. 1996, 18, 1–3.

[8] Rortais, A.; Belyaeva, J.; Gemo, M.; Van Der Goot, E.; Linge, J.P. MedISys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards. Food Res. Int. 2010, 43, 1553–1556.

[9] Bollegala, D., Atanasov, V., Maehara, T., & Kawarabayashi, K. (2018). Classinet–predicting missing features for short-text classification. arXiv:1804.05260.

[10] Bui, D.D.A., & Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. Journal of the American Medical Informatics Association, 21(5), 850–857.

[11] 1 Fahn S, Elton RL, members of UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. Florham Park, NJ: MacMillan Healthcare Information; 1987. p 153–163.

[12] Gromadzka G, Schmidt HH, Genschel J, et al. pH1069Q Mutation in ATP7B and biochemical parameters of copper metabolism and clinical manifestation of Wilson's disease. Mov Disord 2006; 21: 245–248.

[13] Richter-Pechanski P, Geis NA, Kiriakou C et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. Digital Health 2021; 7: 20552076211057662.

[14] Silvestri, S., Islam, S., Papastergiou, S., Tzagkarakis, C., & Ciampi, M. (2023). A machine learning approach for the nlp-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem

[15] Meinert, Edward et al. (2018). Weighing benefits and risks in aspects of security, privacy and adoption of technology in a value-based healthcare system. https://scite.ai/reports/10.1186/s12911- 018-0700-0